What is iClean?

iClean is a web-based tool that enables users to rapidly disambiguate a list of names. Name ambiguity is a problem frequently encountered in published literature where authors may record their names in a variety of ways (e.g., "John Joseph Smith", "John J Smith", and "John Smith" may all refer to the same author).

How does it work?

iClean's name disambiguation routines work by standardizing variations in names within and across different data sources (QVR, PubMed, custom Excel workbooks). When you upload an Excel workbook, you designate which column contains the names that you want to disambiguate. iClean performs the following disambiguation steps:

1. Parses multiple names in a citation string into individual names
2. Standardizes character-level variation (capitalization, accents, hyphens)
3. Identifies suffixes, if any (Jr, Sr, I, II, III, IV)
4. Infers the family and given name from the input, regardless of the original order (e.g., "J Smith" and "Smith, J" both become "{ given: J, family: Smith }")
5. Applies heuristics for determining which names may be merged (e.g., "J Smith" and "John Smith" may merge, but "John Smith" and "Jason Smith" may not)

How is the threshold calculated?

The similarity threshold represents the percentage of character overlap between names and is used to control how closely two names must match in order to be merged. For example, a high similarity threshold would prevent "J Smith" from merging with "John Smith", whereas a lower similarity threshold would not.

How should I prepare my data?

iClean can process data in either Excel (xls/xlsx) or CSV files. The names you wish to have disambiguated will be in a single column with a semicolon separating multiple names. Outputs from QVR (biblio report), SPIRES, Scopus, and Web of Science are all acceptable.

What does the output look like?

The output file will be the same format as the input (i.e. csv input files will output a csv file; an Excel input file will output an Excel file). The output will contain all the data in your input file and will include these extra columns:

- Authors Cleaned: This column contains the cleaned versions of the names and is the column you would use, for example, in constructing a coauthor network.
- Merged Names: This column shows you which names have been merged. Use these data to adjust your threshold if the merging is too stringent or too relaxed. The rows in this column do not correspond to the data in the rows in the rest of the sheet; rather this is a stand-alone list.
- Similarity Threshold: This shows the value of the threshold you chose.

What browsers are supported?

We seek to support all modern browsers and platforms. If you experience a problem using any of the following, please let us know:

- Chrome: all current versions
- Firefox: all current versions
- Internet Explorer: 10 or later
- Opera: all current versions
- Safari: all current versions

How can I send feedback?

Send an email to iClean@mail.nih.gov.